**Linear Regression Assumptions**

The assumptions for linear regression are:

* There is a linear relationship between the dependent variable and the independent variable(s).
* The residuals are normally distributed.
* The variance of the residuals is constant over all values of the independent variable(s) (homoscedasticity).
* The residuals are independent.

These assumptions are important because they ensure that the results of the regression analysis are valid. If the assumptions are not met, the results of the analysis may be misleading.

**Logistic Regression Assumptions**

Assumptions for Binary Logistic Regression

* Binary Outcome: The dependent variable should be binary (i.e., having only two possible outcomes).
* Linearity of Independent Variables and Log Odds: There should be a linear relationship between the independent variables and the log odds of the dependent variable.
* No Multicollinearity: The independent variables should not be highly correlated with each other.
* Large Sample Size: Logistic regression generally requires a larger sample size compared to linear regression for reliable results.
* Independence of Observations: The observations in the dataset should be independent of each other.

It's important to note that logistic regression is more robust to violations of some of these assumptions compared to linear regression. However, checking for and addressing potential violations can improve the accuracy and reliability of your results.

**Terms**

**General Linear Model**

* A statistical model that assumes a linear relationship between the dependent variable and one or more independent variables.
* The basic form is y=β0​+β1​X+e where:  
  + y is the dependent variable.
  + X is the independent variable.
  + β0​ is the intercept.
  + β1​ is the slope.
  + e is the error term or residual.

**Residuals**

* The differences between the actual and predicted values of the dependent variable in a regression model.

**Normally Distributed**

* A probability distribution that is symmetric and bell-shaped.

**Homoscedasticity**

* The variance of the residuals is constant across all levels of the independent variable.

**Independent**

* The residuals are not correlated with each other.

**Skewed**

* A distribution that is not symmetric.

**Linearity**

* A linear relationship between the dependent and independent variables.

**Multi-Collinearity**

* Independent variables in a regression model are highly correlated with each other.

**Hypothesis Testing**

* A statistical method used to determine whether there is enough evidence to support a claim about a population parameter.

**Correlation**

* A statistical measure that describes the strength and direction of the linear relationship between two variables.

**Python**

* A high-level, interpreted, general-purpose programming language.

**Pandas**

* A Python library used for data manipulation and analysis.

**Seaborn**

* A Python library used for data visualization.

**Scikit-learn (sklearn)**

* A Python library used for machine learning.

**Train-test Split**

* A technique used to split a dataset into training and testing sets.

**Quantiles**

* Values that divide a distribution into equal-sized groups.

**Theoretical Quantiles**

* Quantiles that are calculated from a theoretical distribution.

**Sample Quantiles**

* Quantiles that are calculated from a sample of data.

**Effective Rank**

* The number of independent variables in a regression model that are actually contributing to the prediction of the dependent variable.

**Pairplot**

* A Seaborn function that creates a grid of scatterplots showing the relationship between all pairs of variables in a dataset.